

Weighted projected networks: mapping hypergraphs to networks

Eduardo López^{1,2}

¹*CABDyN Complexity Centre, Saïd Business School,
University of Oxford, Park End Street,
Oxford OX1 1HP, United Kingdom*

²*Physics Department, Clarendon Laboratory, University of Oxford,
Parks Road, Oxford OX1 3PU, United Kingdom**

(Dated: December 12, 2012)

Abstract

Many natural, technological, and social systems incorporate multiway interactions, yet are characterized and measured on the basis of weighted pairwise interactions. In this article, I propose a family of models in which pairwise interactions originate from multiway interactions, by starting from ensembles of hypergraphs and applying projections that generate ensembles of *weighted projected networks*. I calculate analytically the statistical properties of weighted projected networks, and suggest ways these could be used beyond theoretical studies. Projected weighted networks typically exhibit weight disorder along links even for very simple generating hypergraph ensembles. Also, as the size of a hypergraph changes, a signature of multiway interaction emerges on projected weighted networks that distinguishes them from fundamentally weighted pairwise networks. I find the percolation threshold and size of the largest component for hypergraphs of arbitrary uniform rank, translate the results into projected networks, and show that the transition is second order. This general approach to network formation has the potential to shed new light on our understanding of weighted networks.

PACS numbers: 89.75.Hc, 02.10.Ox, 64.60.ah, 89.65.-s

*Electronic address: eduardo.lopez@sbs.ox.ac.uk

I. INTRODUCTION

Recent years have seen the growth of complex networks theory, a research area concerned with the general theory of systems of interacting elements [1]. Its relevance has been illustrated in a number of problems, such as infectious disease propagation [2], the strength of social ties [3], data routing in technological networks [4], and motifs in biological networks [5]. An underlying driver for the growth of this field has been the increased availability of digitized information, which can be efficiently analyzed to uncover relations between system elements.

A simplifying assumption that is made in networks theory is to characterize interactions as being exclusively pairwise (each interaction represented by a link between two nodes), often with an associated interaction intensity or weight, generating so-called weighted networks. The reason for this approach is that usually the information available for real systems is relatively limited. Despite these limitations, weighted networks have proven very useful, as a number of measurable network quantities have shown their relevance in application. Examples of these quantities are the distribution of node degree (number of links connecting to a node) [6, 7], optimal path lengths between network nodes [8], and node clustering (a measure of loops of length three) [9]. Other properties that depend on specific groups of links (e.g., network communities) have also proven quite useful [10, 11].

There are situations, however, where it is known that interactions extend to groups larger than two (multiway interactions), and one can use such information to create more accurate models, avoiding the possibility of oversimplified or misleading results. Examples of these situations are, for instance, networks of affiliations [12–14], where nodes representing individuals connect to each other by virtue of their membership to a group such as their family or workplace colleagues; another example are folksonomies [15], systems that encode information of triplets of the following three ingredients: objects, descriptors of the objects, and the individuals making the descriptions. Characterizing these examples by avoiding the pairwise simplification should lead to more informative and reliable results. In this article, I attempt to provide a method to statistically study systems of multiway interactions and relate them with usual network analysis.

Researchers focusing on problems of multiway interactions have proposed mechanisms by which weights are generated as a consequence of these interactions [12]. For instance,

in affiliation networks, when two nodes belong simultaneously to multiple groups, a feature called co-membership, it is assumed that their relationship intensity is equal to the number of groups they both belong to. Perhaps the most appealing feature of these ideas is that they provide a unifying principle to the structure of some interacting systems: the presence of a group generates links, and being part of multiple groups generates weights. Surprisingly, these unifying ideas have received limited attention, perhaps because some of the mathematical models that are required are less straightforward than typical networks.

To model multiway interactions, it is appropriate to use hypergraphs, which are generalizations of networks [16]. They are composed of a set of nodes and a set of hyperedges. Each hyperedge is a group of interconnected nodes (a clique), and the hypergraph is the collection of all the hyperedges and isolated nodes; networks are the specialization of hypergraphs in which all hyperedges are cliques each with only two nodes. The size of a hyperedge is called rank. Hypergraphs are called homogeneous when all hyperedges are equally likely to be present, or heterogeneous when each hyperedge has its own (possibly unique) probability to appear. For the examples mentioned above: in a folksonomy, for instance, hyperedges are all of rank three, whereas in affiliation networks, in principle, hyperedges can have different ranks; both examples are likely to be heterogeneous hypergraphs.

The notion of hypergraphs generating weights is equivalent to constructing networks that represent a projection of a hypergraph. In other words, starting from a hypergraph, one can create an associated set of links that form a *projected weighted network*, where each link weight is given by the structure of the hypergraph and a projection rule. This construction suggests some intriguing possibilities: some data that is typically studied as a network may in fact emerge from underlying hypergraphs. If that is the case, it should be possible to construct hypergraph models and accompanying projections that can fit observed data and narrow down its origins.

In this article, I study homogeneous and heterogeneous entropy maximizing hypergraph ensembles of arbitrary uniform rank r and define general projections of hypergraphs that lead to ensembles of projected weighted networks. The properties of different projections are explored, relating them to measurable network quantities, and suggesting ways to choose the appropriate projection. The percolation threshold and size of the largest connected component of hypergraphs of arbitrary uniform rank are also derived by use of the mapping between the Potts model and percolation theory [17], and the results are then translated into

the percolation properties of the projected networks. These results show that the transition is of second order. I find that, as a function of size, there are measurable quantities on projected network ensembles of hypergraphs that represent signatures of the presence of hidden multiway relations: when faced with a weighted network, these signatures could provide indications that there is an associated hypergraph hidden in the data.

The article is structured in the following way: I first focus on the general definitions of projections of hypergraphs onto networks, and on models of entropy maximizing ensembles of hypergraphs. With these results, I then study in greater detail the statistical properties of general projected networks, as well as some concrete examples of projection that are likely to occur in empirical and theoretical studies of this problem. These results suggest how to explore network data for possible signatures of multiway relations. Completing the results, I focus on the percolation properties of hypergraphs and their projected networks, and explore the general notion of sparsity. I finalize the article with some discussion and conclusions.

II. MAXIMUM ENTROPY HYPERGRAPHS AND THE NETWORK PROJECTION

Consider a hypergraph, represented by σ , consisting of a set of nodes $1, \dots, N$, and for each possible hyperedge of r nodes i_1, \dots, i_r , an indicator σ_{i_1, \dots, i_r} equal to 1 if the hyperedge is present and 0 if it is absent; all subindices i_1, \dots, i_r take non-repeated values from the set $\{1, \dots, N\}$. In general, a hypergraph does not require r to be the same for all hyperedges. However, for the sake of simplicity, I focus on single rank (all hyperedges have the same r) undirected hypergraphs, with the indicator σ_{i_1, \dots, i_r} symmetric under permutations of i_1, \dots, i_r (if one is interested in studying combinations of rank, one merely requires the introduction of the proper parameters for this, but the qualitative nature of the problem is the same as that studied here).

The general hypergraph projection onto a network is defined as a function \mathcal{P} applied over hyperedges of σ that produces the adjacency matrix w_{ij} for the projected weighted graph G . Network G is formed by the same node set as σ , and its adjacency matrix is w_{ij} . If a node does not belong to any hyperedge, it is isolated in both σ and G . For given σ , one can define the subset $O_{ij}(\sigma) := \{(i_1, \dots, i_r) | (i_1, \dots, i_r) \in \sigma \wedge i \in \{i_1, \dots, i_r\} \wedge j \in \{i_1, \dots, i_r\}\}$ of its hyperedges that include simultaneously nodes i and j . The kinds of projections studied

here are of the type

$$w_{ij}(G) = \mathcal{P}(|O_{ij}(\boldsymbol{\sigma})|), \quad (1)$$

where $o_{ij} \equiv |O_{ij}(\boldsymbol{\sigma})|$ is the size (cardinality) of $O_{ij}(\boldsymbol{\sigma})$. Thus, the weight of link ij in G only depends on the number of hyperedges that contain i and j , an intuitive choice, although certainly not the only possible model.

On a concrete empirical case, the projection \mathcal{P} should reflect the understanding of the relation between $\boldsymbol{\sigma}$ and G . Here, I present results for some reasonable sample choices of $\mathcal{P}(O_{ij})$, namely $\mathcal{P}_a(O_{ij}) = o_{ij}$ (additive projection) and $\mathcal{P}_n(O_{ij}) = \theta(o_{ij})$ (nominal projection), where θ is the Heaviside step function ($= 0$ if the argument is 0 or less, and 1 otherwise). In addition, I show some features satisfied by the projected networks generated by a large class of projections with the general form of Eq. (1). To perform calculations, note that the additive projection can be written in terms of σ_{i_1, \dots, i_r} as

$$\mathcal{P}_a(O_{ij}(\boldsymbol{\sigma})) = o_{ij} = \sum_{(i_1, \dots, i_r) \in O_{ij}(\boldsymbol{\sigma})} \sigma_{i_1, \dots, i_r}, \quad (2)$$

whereas the nominal projection is represented by

$$\mathcal{P}_n(O_{ij}(\boldsymbol{\sigma})) = \theta \left(\sum_{(i_1, \dots, i_r) \in O_{ij}(\boldsymbol{\sigma})} \sigma_{i_1, \dots, i_r} \right). \quad (3)$$

An illustration of \mathcal{P}_a for the case of $r = 3$ is shown in Fig. 1.

In the literature, both hypergraphs and projections have been used to study interaction data qualitatively embedded in complex networks theory, but without a sense of unification. For instance, the choice $\mathcal{P}_n(O_{ij})$ is implicit in work such as [18]; there, if σ_{i_1, \dots, i_r} is interpreted as a specific motif (structural pattern), the model generates unweighted networks guaranteed to possess those motifs. In another approach, found in Refs. [19, 20], each hypergraph (containing $r = 2$ and 3 only) treats each rank separately in that the interactions of nodes by way of pairs is counted independently to the triplet interactions, with no notion of projection onto a simple graph. References [12, 15] do consider projections in some form, but are limited by rank r of hypergraph and by the nature of the projection. Projection \mathcal{P}_a is in fact common [12], and it is often used as a way to characterize the one-mode networks that emerge from bipartite graphs [21]. Equation (1) offers a different way to relate graphs and hypergraphs, which can be applied in the previous models to develop additional understanding of the problems.

To build unbiased statistical models, I adapt the canonical ensemble approach developed in Ref. [22] to hypergraphs. The set of all possible hypergraphs σ is given by $\{\sigma\}_{\text{conf}}$ (the ensemble), or in other words, $\{\sigma\}_{\text{conf}}$ is the union of all possible unique hypergraphs σ . To analytically formulate the ensemble problem, consider the entropy S , defined as

$$S = - \sum_{\{\sigma\}_{\text{conf}}} P(\sigma) \ln P(\sigma), \quad (4)$$

where $P(\sigma)$ represents the probability of a given configuration within the hypergraph ensemble, and the sum over configurations is equivalent to summing over all hyperedge combinations, or $\sum_{\{\sigma\}_{\text{conf}}} \rightarrow \sum_{\sigma_1, \dots, r=0}^1 \cdots \sum_{\sigma_{N-r+1}, \dots, N=0}^1$. The canonical ensemble approach finds the distribution $P(\sigma)$ that maximizes S while satisfying conditions that define the ensemble of interest. Such conditions, say $\{\langle X_\alpha \rangle\}$, with α an enumeration index, are taken to be of the form

$$\sum_{\{\sigma\}_{\text{conf}}} X_\alpha(\sigma) P(\sigma) = \langle X_\alpha \rangle. \quad (5)$$

Finally, since $P(\sigma)$ are probabilities, one must guarantee normalization, which translates into

$$\sum_{\{\sigma\}_{\text{conf}}} P(\sigma) = 1. \quad (6)$$

The solution to this problem ($P(\sigma)$ satisfying the conditions above) is obtained via Lagrange multipliers. Each condition is related to a multiplier, and one solves the equations

$$\frac{\partial}{\partial P(\sigma)} \left[S + \eta \left(1 - \sum_{\{\sigma\}_{\text{conf}}} P(\sigma) \right) + \sum_{\alpha} \beta_{\alpha} \left(\langle X_{\alpha} \rangle - \sum_{\{\sigma\}_{\text{conf}}} X_{\alpha}(\sigma) P(\sigma) \right) \right] = 0, \quad (7)$$

for $P(\sigma)$, with η, β_1, \dots the Lagrange multipliers. The solution to the problem can be expressed as

$$P(\sigma) = \frac{e^{-H(\sigma)}}{Z}. \quad (8)$$

The partition function Z , and $H(\sigma)$ (*defined* as the Hamiltonian), are respectively given by

$$Z = \sum_{\{\sigma\}_{\text{conf}}} e^{-H(\sigma)} \quad (9)$$

and

$$H(\sigma) = \sum_{\alpha} \beta_{\alpha} X_{\alpha}(\sigma). \quad (10)$$

Among the simplest non-trivial problems one can address is that of the fully random hypergraph with equal probability for any hyperedge to exist. The constraint associated with this example is the requirement that there is a given average number of hyperedges, $\langle L_r \rangle$, over the hypergraph ensemble. Since L_r for a given configuration σ is given by $L_r(\sigma) = \sum_{(i_1, \dots, i_r) \in \sigma} \sigma_{i_1, \dots, i_r}$, the set of constraints reduces to two Lagrange multipliers, one for the normalization, and another parameter, labelled β , for $\langle L_r \rangle$. Introducing this in Eq. (7) generates the Hamiltonian

$$H(\sigma) = \beta \sum_{(i_1, \dots, i_r) \in \sigma} \sigma_{i_1, \dots, i_r} = \beta L_r(\sigma) \quad (11)$$

and the partition function

$$\begin{aligned} Z &= \sum_{\{\sigma\}_{\text{conf}}} e^{-\beta \sum_{(i_1, \dots, i_r) \in \sigma} \sigma_{i_1, \dots, i_r}} \\ &= \sum_{\sigma_1, \dots, r=0}^1 \cdots \sum_{\sigma_{N-r+1}, \dots, N=0}^1 \prod_{(i_1, \dots, i_r) \in \mathbf{T}(N, r)} e^{-\beta \sigma_{i_1, \dots, i_r}} \\ &= \sum_{\sigma_1, \dots, r=0}^1 e^{-\beta \sigma_1, \dots, r} \cdots \sum_{\sigma_{N-r+1}, \dots, N=0}^1 e^{-\beta \sigma_{N-r+1}, \dots, N} = (1 + e^{-\beta})^{\binom{N}{r}}, \quad (12) \end{aligned}$$

where $\mathbf{T}(N, r)$ is the set of all possible hyperedges $\{(1, \dots, r), \dots, (N-r+1, \dots, N)\}$, i.e., the complete hypergraph of single rank r and size N . The last equality can also be obtained from the symmetry of the Hamiltonian over exchange of indices among σ_{i_1, \dots, i_r} . The result expresses that there are $\binom{N}{r}$ possible hyperedges (i_1, \dots, i_r) among the N nodes. Using this result one can show that the $\langle L_r \rangle$ constraint is satisfied for the proper choice of β , as seen from averaging $L_r(\sigma)$ in the $P(\sigma)$ ensemble

$$\langle L_r \rangle = \frac{1}{Z} \sum_{\{\sigma\}_{\text{conf}}} \sum_{(i_1, \dots, i_r) \in \sigma} \sigma_{i_1, \dots, i_r} e^{-\beta \sum_{(j_1, \dots, j_r) \in \sigma} \sigma_{j_1, \dots, j_r}} = \frac{1}{Z} \sum_{L_r=0}^{\binom{N}{r}} \binom{\binom{N}{r}}{L_r} L_r e^{-\beta L_r} = \binom{N}{r} p, \quad (13)$$

and $p \equiv (1 + e^\beta)^{-1}$ is the probability for a hyperedge to be present, which is evident from writing $\langle L_r \rangle / \binom{N}{r} = (1 + e^\beta)^{-1} = p$; p also corresponds to the expectation value of any hyperedge $\langle \sigma_{i_1, \dots, i_r} \rangle = \sum_{\{\sigma\}_{\text{conf}}} \sigma_{i_1, \dots, i_r} P(\sigma) = (1 + e^\beta)^{-1} = p$, i.e., the probability for any hyperedge to exist. The fact that all hyperedges are equally likely suggests referring to this case as the *homogeneous hypergraph ensemble*. The probability of a specific hypergraph configuration to be observed is given by Eq. (8), which in this case yields

$$P(\sigma, p) = p^{L_r(\sigma)} (1 - p)^{\binom{N}{r} - L_r(\sigma)} \quad (14)$$

where the relations $1 + e^{-\beta} = (1 - p)^{-1}$ and $e^{-\beta} = p(1 - p)^{-1}$ have been used. The application of the \mathcal{P}_a and \mathcal{P}_n to the homogeneous ensemble is tackled below in a more general ensemble.

The solution to the simple homogeneous problem above, helps to identify some basic features of the canonical approach, including quantities such as the probability of a hyperedge, and of a specific hypergraph σ . Building on this, we can construct the more general *heterogeneous* case, where each hyperedge has its own expectation $\langle \sigma_{i_1, \dots, i_r} \rangle$. Thus, the hamiltonian of Eq. (10) becomes

$$H(\sigma) = \sum_{(i_1, \dots, i_r) \in \sigma} \beta_{i_1, \dots, i_r} \sigma_{i_1, \dots, i_r}. \quad (15)$$

In analogy with the homogeneous case, one defines $p_{i_1, \dots, i_r} \equiv \langle \sigma_{i_1, \dots, i_r} \rangle = (1 + e^{\beta_{i_1, \dots, i_r}})^{-1}$. The partition function becomes

$$Z(\mathbf{p}) = \prod_{(i_1, \dots, i_r) \in \mathbf{T}(N, r)} (1 + e^{-\beta_{i_1, \dots, i_r}}) = \prod_{(i_1, \dots, i_r) \in \mathbf{T}(N, r)} (1 - p_{i_1, \dots, i_r})^{-1}, \quad (16)$$

where \mathbf{p} represents the hyperedge expectations $\{p_{1, \dots, r}, \dots, p_{N-r+1, \dots, N}\}$. The probability of a hypergraph configuration σ is then

$$P(\sigma, \mathbf{p}) = \prod_{(i_1, \dots, i_r) \in \mathbf{T}(N, r)} p_{i_1, \dots, i_r}^{\sigma_{i_1, \dots, i_r}} (1 - p_{i_1, \dots, i_r})^{1 - \sigma_{i_1, \dots, i_r}}, \quad (17)$$

which is the joint probability that hyperedges with $\sigma_{i_1, \dots, i_r} = 1$ are present, and those with $\sigma_{i_1, \dots, i_r} = 0$ are absent. If for all $(i_1, \dots, i_r) \in \mathbf{T}(N, r)$, $p_{i_1, \dots, i_r} = p$, one recovers the homogeneous case. The heterogeneous ensemble possesses the most degrees of freedom among non-interacting undirected hypergraph models. If more specific constraints are imposed such as, for instance, conditions on the average number of hyperedges visiting a node, they would translate into additional constraints on the values of the set \mathbf{p} .

III. APPLICATION OF THE HYPERGRAPH PROJECTION

Since only projections of the form $\mathcal{P}(O_{ij}) = \mathcal{P}(o_{ij})$ are considered here, the statistical properties of the projected networks depend on the statistical properties of o_{ij} . It is most useful to focus on the distribution of o_{ij} in the heterogeneous ensemble, $\phi_{ij}(o_{ij}, \mathbf{p})$, and determine how this translates into the homogeneous case (Table I summarizes the notation used to compute $\phi_{ij}(o_{ij}, \mathbf{p})$). Let us define $\mathbf{T}_{ij}(N, r)$ as the set of all hyperedges on the complete hypergraph that visit i and j simultaneously. We also define $\overline{\mathbf{T}}_{ij}(N, r)$, the complement of

$\mathbf{T}_{ij}(N, r)$ with respect to $\mathbf{T}(N, r)$. In addition, for each configuration σ , $V_{ij}(\sigma)$ is the set of hyperedges visiting i and j which may or may not have cardinality o_{ij} (if it does, it is represented as before with $O_{ij}(\sigma)$). The complement of $V_{ij}(\sigma)$ with respect to $\mathbf{T}_{ij}(N, r)$ is $\overline{V}_{ij}(\sigma)$, and thus $\mathbf{T}_{ij}(N, r) = V_{ij}(\sigma) \cup \overline{V}_{ij}(\sigma)$. Taking this into account, $\phi_{ij}(o_{ij}, \mathbf{p})$ can be calculated through the expression

$$\phi_{ij}(o_{ij}, \mathbf{p}) = \sum_{\{\sigma\}_{\text{conf}}} \delta \left(o_{ij}, \sum_{(i_1, \dots, i_r) \in V_{ij}(\sigma)} \sigma_{i_1, \dots, i_r} \right) P(\sigma, \mathbf{p}), \quad (18)$$

where $\delta(x, y, \dots, z)$ is the Kronecker delta which can have two or more arguments, and is equal to 1 if all the arguments are equal, and 0 otherwise. In the sum above, only those configurations for which delta is 1 contribute to $\phi_{ij}(o_{ij}, \mathbf{p})$, and this occurs only when there are exactly o_{ij} hyperedges in σ that include ij .

To perform the calculation, note the independence of each component of \mathbf{p} in Eq. (17). This allows factoring the sum over configurations in Eq. (18) into a product of i) the configurations of hyperedges $\overline{\mathbf{T}}_{ij}(N, r)$, which cannot affect the delta, and ii) the configurations of hyperedges $\mathbf{T}_{ij}(N, r)$, which can. The hyperedges $(i_1, \dots, i_r) \in \overline{\mathbf{T}}_{ij}(N, r)$ each contribute a factor $\sum_{\sigma_{i_1, \dots, i_r}=0}^1 p_{i_1, \dots, i_r}^{\sigma_{i_1, \dots, i_r}} (1 - p_{i_1, \dots, i_r})^{1-\sigma_{i_1, \dots, i_r}} = 1$. Therefore, the remaining factors of Eq. (18) lead to

$$\begin{aligned} \phi_{ij}(o_{ij}, \mathbf{p}) &= \sum_{V_{ij}(\sigma) \subset \mathbf{V}_{ij}} \delta \left(o_{ij}, \sum_{(i_1, \dots, i_r) \in V_{ij}(\sigma)} \sigma_{i_1, \dots, i_r} \right) \prod_{(i_1, \dots, i_r) \in \overline{\mathbf{T}}_{ij}(N, r)} p_{i_1, \dots, i_r}^{\sigma_{i_1, \dots, i_r}} (1 - p_{i_1, \dots, i_r})^{1-\sigma_{i_1, \dots, i_r}} \\ &= \sum_{V_{ij}(\sigma) \subset \mathbf{V}_{ij}} \delta \left(o_{ij}, \sum_{(i_1, \dots, i_r) \in V_{ij}(\sigma)} \sigma_{i_1, \dots, i_r} \right) \prod_{(i_1, \dots, i_r) \in V_{ij}(\sigma)} p_{i_1, \dots, i_r} \prod_{(i_1, \dots, i_r) \in \overline{V}_{ij}(\sigma)} (1 - p_{i_1, \dots, i_r}) \\ &= \sum_{O_{ij}(\sigma) \subset \mathbf{O}_{ij}} \prod_{(i_1, \dots, i_r) \in O_{ij}(\sigma)} p_{i_1, \dots, i_r} \prod_{(i_1, \dots, i_r) \in \overline{O}_{ij}(\sigma)} (1 - p_{i_1, \dots, i_r}), \quad (19) \end{aligned}$$

where \mathbf{V}_{ij} and \mathbf{O}_{ij} are the unions of all possible sets $V_{ij}(\sigma)$ and $O_{ij}(\sigma)$, respectively, and $\overline{O}_{ij}(\sigma)$ is the complement of $O_{ij}(\sigma)$ with respect to $\mathbf{T}_{ij}(N, r)$.

Equation (19) has been expressed in a way that makes it straightforward to explain and convert into an algorithm for calculation, as I attempt to explain now. The expression can be described in the following terms: i) separate the hyperedges from $\mathbf{T}(N, r)$ into two groups, one that can influence ij over all possible configurations, namely $\mathbf{T}_{ij}(N, r)$, and another that cannot ($\overline{\mathbf{T}}_{ij}(N, r)$), ii) identify out of $\mathbf{T}_{ij}(N, r)$ the hyperedges of σ visiting i and j ,

$V_{ij}(\sigma)$, iii) only when $|V_{ij}(\sigma)| \equiv |O_{ij}(\sigma)| = o_{ij}$, σ contributes to $\phi_{ij}(o_{ij}, \mathbf{p})$, and iv) since there are numerous ways to choose $O_{ij}(\sigma)$ from $\mathbf{T}_{ij}(N, r)$, one requires the set \mathbf{O}_{ij} , which contains all those choices, i.e., is the ensemble of allowed configurations. Consequently, the last line of Eq. (19) can be read as the sum of probabilities over all possible configurations \mathbf{O}_{ij} of hyperedge sets $O_{ij}(\sigma)$, where each hyperedge belongs to $\mathbf{T}_{ij}(N, r)$. Note that there are $|\mathbf{T}_{ij}(N, r)| = \binom{N-2}{r-2}$ hyperedges to choose from and each $O_{ij}(\sigma)$ (configuration σ) picks o_{ij} of them, and therefore $|\mathbf{O}_{ij}| = \binom{N-2}{o_{ij}}$. It is worth mentioning that σ is used in $O_{ij}(\sigma)$ to emphasize its origin as a particular hypergraph configuration, but that it becomes redundant when the meaning of \mathbf{O}_{ij} is fixed as the collection of configurations (the ensemble) contributing to $\phi_{ij}(o_{ij}, \mathbf{p})$; at this point, each O_{ij} specifies a unique configuration and no further reference to σ is necessary.

In fact, dropping σ from O_{ij} offers a combinatorial picture for the last line of Eq. (19) and other distributions in this section. Since each σ is a set of hyperedges connecting non-repeated nodes in cliques of rank r , one can think of each hyperedge as an r -tuple of non-repeated indices taken from $\{1, \dots, N\}$, and a configuration σ as a collection of non-repeated r -tuples. Therefore, $\mathbf{T}(N, r)$ is the collection of all possible r -tuples, $\mathbf{T}_{ij}(N, r)$ the subset of $\mathbf{T}(N, r)$ containing all r -tuples that have indices i and j simultaneously, each O_{ij} a sample without replacement of o_{ij} r -tuples taken from $\mathbf{T}_{ij}(N, r)$, and \mathbf{O}_{ij} the collection of all possible samplings. This way to think about Eq. (19) transfer the emphasis from a graph theoretic problem to a purely combinatorial one. The cardinalities of all the sets calculated before follow naturally.

The average of o_{ij} can be determined from Eq. (19), through $\langle o_{ij}(\mathbf{p}) \rangle = \sum_{o_{ij}=0}^{\binom{N-2}{r-2}} o_{ij} \phi(o_{ij}, \mathbf{p})$, or by calculating $\sum_{\{\sigma\}_{\text{conf}}} o_{ij}(\sigma) P(\sigma)$. The result is

$$\langle o_{ij}(\mathbf{p}) \rangle = \sum_{(i_1, \dots, i_r) \in \mathbf{T}_{ij}(N, r)} p_{i_1, \dots, i_r} \quad (20)$$

which fits intuition, stating that the expectation of the number of hyperedges visiting the pair ij , is the sum of expectations of each hyperedge that can visit ij to be present over the ensemble.

In the homogeneous case, where $p_{i_1, \dots, i_r} = p$ for all (i_1, \dots, i_r) , $\phi_{ij}(o_{ij}, \mathbf{p}) \rightarrow \phi_{ij}(o_{ij}, p)$ becomes

$$\phi_{ij}(o_{ij}, p) = \binom{\binom{N-2}{r-2}}{o_{ij}} p^{o_{ij}} (1-p)^{\binom{N-2}{r-2} - o_{ij}}, \quad (21)$$

a binomial distribution with $\langle o_{ij} \rangle = \binom{N-2}{r-2} p$ (Fig. 2(a)). This average has an interesting interpretation explained below regarding what sorts of signatures a multiway interaction may provide in observational studies. Another noteworthy fact stemming from Eq. (21), even in this very simple case of homogeneous p , is that links display disorder in o_{ij} , and this could easily pass on to the projected networks in many kinds of projections. Further structure can be given to this disorder by changing the projection and/or the hypergraph ensemble.

Some general features of \mathcal{P} can now be described. First, note that monotonic smooth projections, satisfying the inverse function theorem, offer a way to formally write the distribution of w_{ij} from the distribution of o_{ij} because there is a one-to-one relation between the two quantities. Defining the distribution $\mu_{ij}(w_{ij}, \mathbf{p})$, the change of variables theorem for probability distributions implies

$$\mu_{ij}(w_{ij}, \mathbf{p}) = \frac{\phi_{ij}(\mathcal{P}^{-1}(w_{ij}), \mathbf{p})}{\mathcal{P}'(\mathcal{P}^{-1}(w_{ij}))}, \quad (22)$$

where \mathcal{P}' is the derivative of \mathcal{P} . The additive projection \mathcal{P}_a satisfies such conditions in a trivial way because it is just the identity function. However, a large class of functions also satisfy these conditions, including all power law and logarithmic growth or decay functions (when decay applies, one must be mindful of additional conditions). The nominal projection, on the other hand, does not satisfy the condition because any value of $o_{ij} \geq 1$ leads to the same weight $w_{ij} = 1$, and thus the inverse of \mathcal{P}_n is not defined.

An important feature of these mappings is patent in Eq. (21): the distribution of o_{ij} is narrow, with relative width decaying as $\binom{N-2}{r-2}^{-1/2}$, so as N grows, more of the mass of the distribution is concentrated around its maximum, labelled o_{ij}^* , which coincides with the average $\langle o_{ij} \rangle = \binom{N-2}{r-2} p$ for large N . It is then expected that for a wide range of possible projections, asymptotic estimates of $\mu_{ij}(w_{ij}, p)$ can be straightforwardly obtained.

An interesting observation emerges from the previous results. Equation (21) predicts via its average that in projected networks of homogeneous hypergraphs the interaction weight between nodes in the system increases with $\binom{N-2}{r-2}$ or roughly N^{r-2} for large N and finite $r > 2$ (but $N \gg r$). This also indicates the following: since nodes added to the hypergraph only establish interactions with other nodes by way of hyperedges, the addition of these nodes increases on average the interaction weights among *all* nodes, i.e., new and already present. In contrast, with only pairwise interactions, the addition of new nodes would *not*

contribute to the weight of interactions of nodes that are already present, because the new nodes require only direct new connections to existing nodes one at a time, and thus do not effect changes in the weights between existing nodes. This mechanism distinguishes networks that are fundamentally pairwise in origin to those which may *appear* as pairwise due to data collection or other factors but are, in fact, fundamentally hypergraphs. This observation may suggest tests to distinguish the two scenarios.

The two projections \mathcal{P}_a and \mathcal{P}_n can now be explained further. For \mathcal{P}_a , the properties of w_{ij} are those of o_{ij} , and thus already calculated. The other property to describe is the so-called strength s_i of node i , equal to $\sum_j o_{ij}$. It is intuitively helpful to calculate the distribution of strengths $\xi_i(s_i, \mathbf{p})$ by making use of the relation between s_i and ℓ_i , the number of hyperedges visiting i . These two quantities relate via $s_i = (r-1)\ell_i$, and one can determine the distribution $\zeta_i(\ell_i, \mathbf{p})$ of ℓ_i and from it compute $\xi_i(s_i, \mathbf{p})$. Note that while s_i is a property of the graph, ℓ_i is a property of the hypergraph. Once again, the independence of the components of \mathbf{p} simplifies the sum over configurations $\{\boldsymbol{\sigma}\}_{\text{conf}}$ (notation in Table II). The hyperedges that could affect ℓ_i belong to $\mathbf{T}_i(N, r)$, the collection of all hyperedges visiting i in $\mathbf{T}(N, r)$, and $L_i(\boldsymbol{\sigma})$ is the set of hyperedges from $\mathbf{T}_i(N, r)$ in configuration $\boldsymbol{\sigma}$ (when $|L_i(\boldsymbol{\sigma})| = \ell_i$ we write it as $\lambda_i(\boldsymbol{\sigma})$). From the definition $\zeta_i(\ell_i, \mathbf{p}) = \sum_{\{\boldsymbol{\sigma}\}_{\text{conf}}} \delta(\ell_i, \sum_{(i_1, \dots, i_r) \in L_i(\boldsymbol{\sigma})} \sigma_{i_1, \dots, i_r}) P(\boldsymbol{\sigma}, \mathbf{p})$, one can quickly conclude that

$$\zeta_i(\ell_i, \mathbf{p}) = \sum_{\lambda_i(\boldsymbol{\sigma}) \in \Lambda_i} \prod_{(i_1, \dots, i_r) \in \lambda_i(\boldsymbol{\sigma})} p_{i_1, \dots, i_r} \prod_{(i_1, \dots, i_r) \in \bar{\lambda}_i(\boldsymbol{\sigma})} (1 - p_{i_1, \dots, i_r}), \quad (23)$$

where Λ_i is the ensemble of configurations $\lambda_i(\boldsymbol{\sigma})$, and $\lambda_i(\boldsymbol{\sigma}) \cup \bar{\lambda}_i(\boldsymbol{\sigma}) = \mathbf{T}_i(N, r)$. Then

$$\xi_i(s_i, \mathbf{p}) = \zeta_i(s_i/(r-1), \mathbf{p})/(r-1), \quad (24)$$

where s_i takes values $0, r-1, 2(r-1), \dots, (r-1)\binom{N-1}{r-1}$. Once again, an equivalence between hyperedge sets and combinatorics can be drawn: $\mathbf{T}_i(N, r)$ is the union of all r -tuples drawn from $\{1, \dots, N\}$ with one element always i , and thus there are $|\mathbf{T}_i(N, r)| = \binom{N-1}{r-1}$ r -tuples in total. Each λ_i is a distinct choice of ℓ_i of these r -tuples; clearly $|\Lambda_i| = \binom{N-1}{\ell_i}$. The sum $\sum_{\lambda_i \in \Lambda_i}$ is a sum over all choices of ℓ_i r -tuples from $\mathbf{T}_i(N, r)$. The averages of these quantities are given by

$$\langle \ell_i(\mathbf{p}) \rangle = \sum_{(i_1, \dots, i_r) \in \mathbf{T}_i(N, r)} p_{i_1, \dots, i_r} \quad (25)$$

and

$$\langle s_i(\mathbf{p}) \rangle = (r-1)\langle \ell_i(\mathbf{p}) \rangle = (r-1) \sum_{(i_1, \dots, i_r) \in \mathbf{T}_i(N, r)} p_{i_1, \dots, i_r}. \quad (26)$$

For the homogeneous case, $\zeta_i(\ell_i, p) = \binom{N-1}{r-1} p^{\ell_i} (1-p)^{\binom{N-1}{r-1} - \ell_i}$, with average $\langle \ell_i \rangle = \binom{N-1}{r-1} p$.

Therefore,

$$\xi_i(s_i, p) = \binom{N-1}{s_i/(r-1)} p^{s_i/(r-1)} (1-p)^{\binom{N-1}{r-1} - s_i/(r-1)}, \quad (27)$$

and $\langle s_i \rangle = (r-1) \binom{N-1}{r-1} p$ (see Fig. 2(b)).

The nominal interaction \mathcal{P}_n needs a different treatment. Note that under this projection, w_{ij} can be either 0 or 1. To determine the probability for w_{ij} , $\pi_{ij}(w_{ij}, \mathbf{p})$, one merely needs to determine the probabilities that o_{ij} is either 0 or ≥ 1 , that is $\pi_{ij}(w_{ij}, \mathbf{p}) = \phi_{ij}(o_{ij} = 0, \mathbf{p})$ or $\pi_{ij}(w_{ij} = 1, \mathbf{p}) = 1 - \phi_{ij}(o_{ij} = 0, \mathbf{p})$. Therefore,

$$\pi_{ij}(w_{ij}, \mathbf{p}) = \begin{cases} 1 - \prod_{(i_1, \dots, i_r) \in \mathbf{T}_{ij}(N, r)} (1 - p_{i_1, \dots, i_r}); & w_{ij} = 1 \\ \prod_{(i_1, \dots, i_r) \in \mathbf{T}_{ij}(N, r)} (1 - p_{i_1, \dots, i_r}); & w_{ij} = 0. \end{cases} \quad (28)$$

In the homogeneous case,

$$\pi_{ij}(w_{ij}, p) = \begin{cases} 1 - (1-p)^{\binom{N-2}{r-2}}; & w_{ij} = 1 \\ (1-p)^{\binom{N-2}{r-2}}; & w_{ij} = 0. \end{cases} \quad (29)$$

These results have implications for the average number of connections for each node of a projected network, as we explain next.

For network projections \mathcal{P}_a and \mathcal{P}_n , the number of connections k_i visiting node i are characterized by $\psi_i(k_i, \mathbf{p})$, the distribution of k_i . The degree can be either 0 or take any value from $r-1 \leq k_i \leq N-1$. To determine $\psi_i(k_i, \mathbf{p})$ (notation in Table III), one can proceed in a similar way as before: in configuration σ , the set of hyperedges visiting i and producing degree k_i is $K_i(\sigma)$. This means that hyperedges in $K_i(\sigma)$ visit exactly k_i nodes and node i . It is interesting to note that another configuration σ' , associated with $K_i(\sigma')$, with a different set and/or number of hyperedges can lead to the same k_i , because these hyperedges still visit the same number of nodes k_i (see Fig. 3 for an illustration). With this definition, one can write

$$\psi_i(k_i, \mathbf{p}) = \sum_{K_i(\sigma) \in \mathbf{K}_i} \prod_{(i_1, \dots, i_r) \in K_i(\sigma)} p_{i_1, \dots, i_r} \prod_{(i_1, \dots, i_r) \in \overline{K}_i(\sigma)} (1 - p_{i_1, \dots, i_r}), \quad (30)$$

where \mathbf{K}_i is the union of all possible sets $K_i(\sigma)$, and the complement set $\overline{K}_i(\sigma)$ satisfies $K_i(\sigma) \cup \overline{K}_i(\sigma) = \mathbf{T}_i(N, r)$. Since the number of hyperedges is not fixed across members of \mathbf{K}_i , one can further organize the $K_i(\sigma)$ by their numbers of hyperedges $\ell_i(\sigma)$. The bounds of ℓ_i are dictated by the following: for degree k_i , a minimum of $\lceil k_i/(r-1) \rceil$ hyperedges is required ($\lceil \cdot \rceil$ represents the ceiling function), and there can be no more than $\binom{k_i}{r-1}$ hyperedges. Using this organization, and introducing the notation $K_i^{(\ell_i)}(\sigma)$ and $\mathbf{K}_i(\ell_i)$ to represent, respectively, the sets $K_i(\sigma)$ involving exactly ℓ_i hyperedges and their unions, one can write

$$\psi_i(k_i, \mathbf{p}) = \sum_{\ell_i = \lceil \frac{k_i}{r-1} \rceil}^{\binom{k_i}{r-1}} \sum_{K_i^{(\ell_i)}(\sigma) \in \mathbf{K}_i(\ell_i)} \prod_{(i_1, \dots, i_r) \in K_i^{(\ell_i)}(\sigma)} p_{i_1, \dots, i_r} \prod_{(i_1, \dots, i_r) \in \overline{K}_i^{(\ell_i)}(\sigma)} (1 - p_{i_1, \dots, i_r}). \quad (31)$$

The sets $\mathbf{K}_i(\ell_i)$ are only subsets of $\mathbf{\Lambda}_i$ in which the ℓ_i hyperedges involve exactly i and k_i other nodes. Finally, it is possible to exploit one more symmetry that facilitates an algorithmic understanding of $\psi_i(k_i, \mathbf{p})$: the sets that make up $\mathbf{K}_i(\ell_i)$ involve several possible distinct node sets. However, one can further segregate these sets by the specific nodes in them. Hence, if one takes a set, $\rho(k_i)$, of k_i specific nodes and i , there are several configurations in which their associated $K_i^{(\ell_i)}(\sigma)$ contain ℓ_i hyperedges visiting only those nodes. Thus, a configuration with specific $\rho(k_i)$ nodes connected to i , using ℓ_i hyperedges is labelled $I_i^{(\rho(k_i), \ell_i)}(\sigma)$, and the union of configurations is labelled $\mathbf{I}_i(\rho(k_i), \ell_i)$. The union of all sets $\mathbf{I}_i(\rho(k_i), \ell_i)$ (which are non-intersecting) is equal to $\mathbf{K}_i(\ell_i)$. This leads to the final expression

$$\psi_i(k_i, \mathbf{p}) = \sum_{\rho(k_i) \in \mathbf{R}_i(N, k_i)} \sum_{\ell_i = \lceil \frac{k_i}{r-1} \rceil}^{\binom{k_i}{r-1}} \sum_{I_i^{(\rho(k_i), \ell_i)}(\sigma) \in \mathbf{I}_i(\rho(k_i), \ell_i)} \prod_{(i_1, \dots, i_r) \in I_i^{(\rho(k_i), \ell_i)}(\sigma)} p_{i_1, \dots, i_r} \prod_{(i_1, \dots, i_r) \in \overline{I}_i^{(\rho(k_i), \ell_i)}(\sigma)} (1 - p_{i_1, \dots, i_r}), \quad (32)$$

where $\overline{I}_i^{(\rho(k_i), \ell_i)}(\sigma)$ is the complement of $I_i^{(\rho(k_i), \ell_i)}(\sigma)$ with respect to $\mathbf{T}_i(N, r)$, and $\mathbf{R}_i(N, k_i)$ is the union of all possible $\rho(k_i)$, each one a distinct $(k_i + 1)$ -tuple taken from the set $\{1, \dots, N\}$ with one choice always being i . The sizes of sets are: $|\mathbf{R}_i(N, k_i)| = \binom{N-1}{k_i}$, and $|\mathbf{I}_i(\rho(k_i), \ell_i)| = Q_{r-1}(k_i, \ell_i)$; the later is the result of a combinatorial problem that can be defined in terms of general graph theory. Specifically, $Q_{r-1}(k_i, \ell_i)$ corresponds to the number of distinct graphs that can be constructed with k_i nodes all of which belong to at least ℓ_i cliques of size $r - 1$. In fact, each $I_i^{(\rho(k_i), \ell_i)}(\sigma)$ can be mapped to each one of these graphs.

To determine $\langle k_i(\mathbf{p}) \rangle$, it is simpler to use the relation

$$\langle k_i(\mathbf{p}) \rangle = \sum_{\{\boldsymbol{\sigma}\}_{\text{conf}}} k_i(\boldsymbol{\sigma}) P(\boldsymbol{\sigma}) = \sum_{\{\boldsymbol{\sigma}\}_{\text{conf}}} \sum_{j=1; j \neq i}^N \theta(o_{ij}(\boldsymbol{\sigma})) P(\boldsymbol{\sigma}) = \sum_{j=1; j \neq i}^N \sum_{\{\boldsymbol{\sigma}\}_{\text{conf}}} \theta(o_{ij}(\boldsymbol{\sigma})) P(\boldsymbol{\sigma}). \quad (33)$$

By first summing over a single j , one notices that only hyperedges in $\mathbf{T}_{ij}(N, r)$ are relevant, all others contributing a factor of 1, and that

$$\theta(o_{ij}(\boldsymbol{\sigma})) = \theta \left[\sum_{(i_1, \dots, i_r) \in O_{ij}(\boldsymbol{\sigma})} \sigma_{i_1, \dots, i_r} \right] = 1 - \delta \left[\sum_{(i_1, \dots, i_r) \in O_{ij}(\boldsymbol{\sigma})} \sigma_{i_1, \dots, i_r}, 0 \right] \quad (34)$$

one arrives at

$$\langle k_i(\mathbf{p}) \rangle = \sum_{j=1; j \neq i}^N \left[1 - \prod_{(i_1, \dots, i_r) \in \mathbf{T}_{ij}(N, r)} (1 - p_{i_1, \dots, i_r}) \right]. \quad (35)$$

When compared with Eq. (28), it becomes evident that each link ij contributes to k_i independently.

In the homogeneous case, making use of the combinatorial results presented, one obtains (Fig. 2(c))

$$\psi_i(k_i, p) = \binom{N-1}{k_i} \sum_{\ell_i = \lceil \frac{k_i}{r-1} \rceil}^{\binom{k_i}{r-1}} Q_{r-1}(k_i, \ell_i) p^{\ell_i} (1-p)^{\binom{N-1}{r-1} - \ell_i}. \quad (36)$$

Without diving into too much detail, $Q_{r-1}(k_i, \ell_i)$ can be calculated via the inclusion-exclusion principle of combinatorics [25], which produces

$$Q_{r-1}(k_i, \ell_i) = \sum_{m=0}^{k_i} (-1)^{k_i-m} \binom{k_i}{m} \binom{m}{\ell_i}. \quad (37)$$

Among the identities satisfied by $Q_{r-1}(k, \ell)$, one finds that $\binom{\binom{N-1}{r-1}}{\ell} = \sum_k \binom{N-1}{k} Q_{r-1}(k, \ell)$, which is used to show normalization of $\psi(k_1, p)$. Another identify, $\sum_k k \binom{N-1}{k} Q_{r-1}(k, \ell) = (N-1) \left[\binom{\binom{N-1}{r-1}}{\ell} - \binom{\binom{N-2}{r-1}}{\ell} \right]$, leads to the average of $\psi_i(k_i, p)$,

$$\langle k_i \rangle = (N-1) \left[1 - (1-p)^{\binom{N-2}{r-1}} \right], \quad (38)$$

where the brackets are equal to $\pi_{ij}(w_{ij} = 1, p)$ from Eq. (29) (see Fig. 2(d)). This average can also be calculated directly from Eq. (35).

To conclude this section, it is useful to point out how the previous results can be connected with concrete problems. The logic is similar to that found in [22, 23], in which the ensemble

is chosen to fit observations. In the framework presented here, it is possible to choose the hypergraph ensemble to fit hypergraph properties (such as Eqs. (20) or (25)), projected network properties (Eqs. (26) or (35)), or a combination of both (as long as it is well defined); the choice comes down to practical considerations such as the available data one intends to fit, or the belief that certain mechanisms may be at play and therefore must be part of the model. Once an ensemble is defined (satisfying the assumptions of hyperedges which are non-interacting, undirected, and with uniform rank), the expressions derived above for the heterogeneous ensemble apply, but an additional set of constraints emerges for the p_{i_1, \dots, i_r} guaranteeing that the entropy is maximized, distinguishing the situation from that of the fully heterogeneous ensemble, where each p_{i_1, \dots, i_r} is free to have any value between 0 and 1.

As an example, consider the ensemble that specifies strengths $\langle s_i \rangle$ on the projected networks with projection \mathcal{P}_a . This can be constructed from the Hamiltonian

$$H(\boldsymbol{\sigma}) = \sum_{i=1}^N \beta_i s_i(\boldsymbol{\sigma}) = (r-1) \sum_{(i_1, \dots, i_r) \in \boldsymbol{\sigma}} (\beta_{i_1} + \dots + \beta_{i_r}) \sigma_{i_1, \dots, i_r}. \quad (39)$$

This ensemble is completely specified by calculating the relation between $\langle \sigma_{i_1, \dots, i_r} \rangle$, by definition equal to p_{i_1, \dots, i_r} , and the set of parameters $\{\beta_1, \dots, \beta_N\}$. After determining $P(\boldsymbol{\sigma})$, one can compute $\langle \sigma_{i_1, \dots, i_r} \rangle = \sum_{\{\boldsymbol{\sigma}\}_{\text{conf}}} \sigma_{i_1, \dots, i_r} P(\boldsymbol{\sigma})$ to find

$$\langle \sigma_{i_1, \dots, i_r} \rangle = p_{i_1, \dots, i_r} = \frac{e^{-(r-1)(\beta_{i_1} + \dots + \beta_{i_r})}}{1 + e^{-(r-1)(\beta_{i_1} + \dots + \beta_{i_r})}} \quad (40)$$

where the parameters satisfy Eq. (26), and therefore

$$\langle s_i \rangle = (r-1) \sum_{(i_1, \dots, i_r) \in \mathbf{T}_i(N, r)} p_{i_1, \dots, i_r} = (r-1) \sum_{(i_1, \dots, i_r) \in \mathbf{T}_i(N, r)} \frac{e^{-(r-1)(\beta_{i_1} + \dots + \beta_{i_r})}}{1 + e^{-(r-1)(\beta_{i_1} + \dots + \beta_{i_r})}}. \quad (41)$$

One way to understand this result is from the relation

$$dp_{i_1, \dots, i_r} = -(r-1) \sum_{g=1}^r \frac{p_{i_1, \dots, i_r}}{1 - p_{i_1, \dots, i_r}} d\beta_{i_g}. \quad (42)$$

If only β_{i_g} changes by $d\beta_{i_g}$, hyperedges without node i_g are unaffected, and those with it all increase proportionally to $d\beta_{i_g}$. As in Ref. [22], the β_i can be taken from a distribution, leading in turn to a distribution of $\langle s_i \rangle$. This can be used to tune a desired distribution of $\langle s_i \rangle$ as dictated by the problem.

IV. PERCOLATION PROPERTIES AND SPARSE CASES

Another important aspect of the hypergraph ensemble and its projected networks is their percolation properties. To calculate these, one can use the equivalence, first pointed out by Fortuin and Kasteleyn [17], between percolation and the mean-field q -states Potts model at $q \rightarrow 1$. The solution to the later model consists of determining the state of the nodes, and whether there is a phase transition. The solution and its properties can be obtained by studying the model's Helmholtz free energy. A detailed development of equivalence of the models can be found in Ref. [23, 24]; here, I set up the calculation starting at the free energy and develop the percolation properties from there. I consider the homogeneous case only, although it is in principle possible to solve some heterogeneous models.

Consider the Hamiltonian of the general q -state Potts model with N nodes, $H_q = -\sum_{i_1, \dots, i_r} J_{i_1, \dots, i_r} \delta(u_{i_1}, \dots, u_{i_r})$, where u_{i_1}, \dots, u_{i_r} represent the respective spin states of the nodes i_1, \dots, i_r from the possible states $1, \dots, q$, and J_{i_1, \dots, i_r} the strength of the interaction among them. A hyperedge exists among nodes i_1, \dots, i_r if $u_{i_1} = \dots = u_{i_r}$, i.e., if these nodes are in the same spin state. Let us denote the number of system nodes with spin u as N_u , and the density of these as $c_u = N_u/N$, which satisfies $\sum_u c_u = 1$. In the homogeneous system, since $J_{i_1, \dots, i_r} = J$ and given that only r -tuples of equal spins contribute to H_q (i.e., only hyperedges), the energy is equal to $H_q = -J \sum_u \binom{N_u}{r}$, the sum of interaction energies among all hyperedges having equal spin. The connection between percolation and the Potts model translates into the relation $J = -\ln(1 - p)$, and for small p , this approximates to $J \approx p$.

In order to find the Helmholtz free energy of the system, one must first determine the partition function Z_q . In this model, it can be written on the basis of all configurations of state values u_i , or in terms of the densities $\{N_u\}_{u=1, \dots, q}$. Using the later set of variables, and taking into account the multiplicity in the choices for each node state, one arrives at

$$Z_q = \sum_{\sum_{u=1}^q N_u = N} e^{-J \sum_u \binom{N_u}{r}} \frac{N!}{\prod_{u=1}^q N_u!} = \sum_{\sum_{u=1}^q c_u = 1} e^{-[J \sum_u \binom{N c_u}{r} + N \sum_u c_u \ln c_u]}, \quad (43)$$

where the inverse temperature parameter β is absorbed into J . In the canonical ensemble, the free energy is given by $F_q = -\ln Z_q$. When the interaction J is too weak to keep the nodes ordered collectively in groups of common states, the solution to the problem is expected to be symmetric, i.e. $c_u = 1/q$ (all states are equally occupied). However, as

the interaction strengthens, one would expect that symmetry is broken and one state (say $u = 1$) becomes dominant. By these arguments, F_q can be sought by introducing the *ansatz*

$$c_u = \begin{cases} \frac{1+(q-1)\tilde{f}_q}{q} & ; u = 1 \\ \frac{1-\tilde{f}_q}{q} & ; u \neq 1, \end{cases} \quad (44)$$

where \tilde{f}_q is the fractional size of the system in state $u = 1$, and the condition $\sum_u c_u = 1$ is automatically satisfied. This leads to

$$Z_q = \int d\tilde{f}_q e^{\left[J \binom{\frac{N}{q}(1+(q-1)\tilde{f}_q)}{r} + J(q-1) \binom{\frac{N}{q}(1-\tilde{f}_q)}{r} - \frac{N}{q}(1+(q-1)\tilde{f}_q) \ln \frac{1+(q-1)\tilde{f}_q}{q} - \frac{N}{q}(1-\tilde{f}_q) \ln \frac{1-\tilde{f}_q}{q} \right]}. \quad (45)$$

In the thermodynamic limit ($N \rightarrow \infty$), the Laplace method of integration can be applied to Z_q [26]. Once applied, $F_q = -\ln Z_q$ yields to leading order

$$F_q = -J \binom{\frac{N}{q}(1+(q-1)f_q)}{r} - J(q-1) \binom{\frac{N}{q}(1-f_q)}{r} + \frac{N}{q}(1+(q-1)f_q) \ln \frac{1+(q-1)f_q}{q} + \frac{N}{q}(1-f_q) \ln \frac{1-f_q}{q} \quad (46)$$

where f_q is the value of \tilde{f}_q for which the exponent of the argument of the Z_q integral is maximized. Taking the first derivative of the exponent, and using $c_1(f_q)$ and $c_u(f_q)$ to refer to the the fractions c_u from Eq. (44) evaluated at $u = 1$ and $u \neq 1$, respectively, f_q must satisfy

$$\ln \left(\frac{1-f_q}{1-(q-1)f_q} \right) = -J \left[\binom{Nc_1(f_q)}{r} \sum_{i=1}^{r-1} \frac{1}{Nc_1(f_q) - i} - \binom{Nc_u(f_q)}{r} \sum_{i=1}^{r-1} \frac{1}{Nc_u(f_q) - i} \right]. \quad (47)$$

This is the self-consistency equation for the fractional size of the component of broken symmetry. For $q = 1$, $f \equiv f_{q=1}$ is the fractional size of the percolating spanning cluster. Note that $f_q = 0$ is also a solution to Eq. (47), but its stability breaks down when the second derivative of the exponent integrand of Z_q changes sign. This leads to the relation

$$p_c \approx J_c = \left[N \frac{\partial^2}{\partial N^2} \binom{N}{r} \right]^{-1}, \quad (48)$$

where $q = 1$ has already been introduced (otherwise the solution would be the same but with N/q in place of N everywhere).

In the thermodynamic limit, one can derive a compact equation for s and arbitrary r . Both terms in the brackets of Eq. (47) are polynomials emerging from the derivative

$\frac{\partial}{\partial x} \binom{x}{r}$, which we label $M(x, r) = \sum_{m=0}^{r-1} a_r(m) x^m$, evaluated at $x = N_{c_1}$ and N_{c_u} (we continue the same shorthand of $c_{u \neq 1} = c_u$). Subtracting them, one obtains $M(N_1, r) - M(N_{u \neq 1}, r) = \sum_{m=1}^{r-1} a_r(m) (N_1^m - N_u^m)$, where the coefficient $m = 0$ vanishes. It is possible to express the coefficients $a_r(m)$ in terms of elementary symmetric polynomials and binomial coefficients, but the analysis here is restricted to the asymptotic limit, and thus only requires the coefficient $a_r(r-1)$, equal to $1/(r-1)!$ as can be determined by inspecting $M(x, r)$. Using the identity $x^m - y^m = (x-y) \sum_{l=0}^{m-1} x^{m-1-l} y^l$, the ansatz (44), and the self-consistency relation (47) with $q = 1$, one obtains

$$\ln(1-f) = -Jf \sum_{m=1}^{r-1} a_r(m) N^m \sum_{l=0}^{m-1} (1-f)^l = -J \sum_{m=1}^{r-1} a_r(m) N^m (1 - (1-f)^m). \quad (49)$$

Close to percolation, it is justified to write $J = \lambda J_c$, with $\lambda \geq 1$ and J_c from Eq. (48). By L'Hopital's rule, for the dominant term in N , the size of the largest component emerges as

$$\ln(1-f) = -\frac{\lambda}{r-1} (1 - (1-f)^{r-1}), \quad (50)$$

which generalizes expressions for this quantity for specific small r values found in [19, 20]. To test this expression, it is customary to define the percolation problem with respect to a network (or hypergraph) that is not complete, but instead is already diluted. By defining the rescaling $z = p/p_{\max}$ where typically $p_{\max} \ll 1$, the original undiluted hypergraph is $z = 1$, and percolation occurs at $z_c = p_c/p_{\max}$, or if using λ_{\max} , $z_c = 1/\lambda_{\max}$ (see Fig 4(a)).

The percolation transition can be shown to be second order by expanding both sides of Eq. (50), which leads to

$$\sum_{g=1}^{\infty} \frac{f^g}{g} = \frac{\lambda}{r-1} \sum_{g'=1}^{r-1} \binom{r-1}{g'} (-1)^{g'+1} f^{g'}. \quad (51)$$

For small f , close to the percolation transition, only the first few terms on both sides of the equality are relevant. Retaining up to second order

$$1 + \frac{f}{2} \approx \frac{\lambda}{r-1} \left[(r-1) - \binom{r-1}{2} f \right] \quad (52)$$

which produces

$$f \sim \frac{2(\lambda-1)}{1 + (r-2)\lambda} \sim 2(\lambda-1) \quad (53)$$

clearly indicating a continuous transition, in the same universality class of regular network percolation, which diverges at the transition with exponent 1. This result is known in the literature [19].

The previous results focus on hypergraphs, but their relevance to projected networks is not explicitly clear. To clarify this, it is sufficient to explore the properties of $\phi_{ij}(o_{ij}, p)$. For this, it is useful to have in mind the asymptotic relations $N \frac{\partial^2}{\partial N^2} \binom{N}{r} \sim \frac{N^{r-1}}{(r-2)!}$ and $\binom{N-1}{r-1} \sim \frac{N^{r-1}}{(r-1)!}$. Inserting $p = \lambda p_c$ in Eq. (21), and taking the limit $N \gg o_{ij}$, the relation $\phi_{ij, \text{sparse}}(o_{ij}, \lambda p_c) \equiv \phi_{ij}^{(s)}(o_{ij}, \lambda) = (\lambda/N)^{o_{ij}} e^{\lambda/N} / o_{ij}!$, which is a poisson distribution with average λ/N (Fig. 4(b)). Therefore, as N increases, the weights on the edges vanish, signalling the fact that in this dilute regime, the hypergraph and projected networks are virtually the same, and hyperedges are non-overlapping asymptotically. Thus, one only needs to calculate the hypergraph percolation properties to be able to write down the projected network percolation properties. In this sparse regime, the other distributions discussed above have particular forms: for the hypergraph, the distribution of hyperedges visiting a node becomes $\zeta_i^{(s)}(\ell_i, \lambda) = (\lambda/(r-1))^{\ell_i} e^{-\lambda/(r-1)} / \ell_i!$ (poisson with average $\lambda/(r-1)$), and the strength distribution on projected networks with \mathcal{P}_a becomes $\xi_i^{(s)}(s_i, \lambda) = (\lambda/(r-1))^{s_i/(r-1)} e^{-\lambda/(r-1)} / [(r-1)(s_i/(r-1))!]$ (Fig. 4(c)). From these results, the meaning of λ emerges as the parameter that measures the average node strength of the projected network. Finally, the degree distribution can be calculated if one keeps in mind that in the sparse limit, the probability that hyperedges overlap is minimal, and therefore, one expects that only the minimum number of hyperedges $\ell_i \rightarrow \lceil k_i/(r-1) \rceil$ contribute to the distribution. There are subtleties present in explicitly calculating $Q_{r-1}(k_i, \lceil k_i/(r-1) \rceil)$ and $\psi_i^{(s)}(k_i, \lambda)$ when k_i is not a multiple of $r-1$ because hyperedges are forced to overlap in this case, and thus to avoid further details, I only write the unevaluated result $\psi_i^{(s)}(k_i, \lambda) = Q_{r-1}(k_i, \lceil k_i/(r-1) \rceil) (\lambda p_c)^{\lceil k_i/(r-1) \rceil} (1 - \lambda p_c)^{\binom{N-1}{r-1} - \lceil k_i/(r-1) \rceil}$ (Fig. 4(d)). However, the calculations are not prohibitive, and will be derived in detail in a forthcoming publication.

The sparse regime close to percolation is not the only possible sparse regime. To be concrete, note that for p close to p_c , the average node strength is constant, but the average overlap on projected edges vanishes linearly with N , so the larger the network, the less interaction present along the edges. However, one can consider a regime in which $\langle o_{ij} \rangle \sim \lambda/N$ is constant, and in this regime node strength increases with N . Both of these regimes are “sparse” in the sense that p vanishes asymptotically, but each regime has specific properties. Generally, these sparse regimes can be defined based on any sensible property, and lead to interesting behaviors. Finally, for the dense regime (p constant), the interesting effect of

growth of $\langle o_{ij} \rangle$ vs. N emerges, which is a unique feature of this model.

In conclusion, in this article I present a model of hypergraphs and associated projected weighted networks that offers a concise and intuitive picture of hypergraphs, networks, and weights. By using statistical mechanics concepts, together with combinatorial tools, I have been able to determine some basic features of homogeneous and heterogeneous projected networks that offer concrete tests to determine whether a network that has been empirically measured may bear the signature of multiway (group) interactions. The general idea of using the projection of a hypergraph onto a network has not been well studied, and deserves a close look to determine further properties that can help give a better understanding to the genuine limits and virtues of pairwise network analysis.

The author thanks L. Roberts, A. Gerig, F. Reed-Tsochas, and O. Riordan, for helpful discussions, and TSB/EPSRC grant SATURN (TS/H001832/1) and ICT eCollective EU project (238597) for financial support.

-
- [1] R. Albert and A.-L. Barabási, *Rev. Mod. Phys.* **74**, 47 (2002); R. Pastor-Satorras and A. Vespignani, *Structure and Evolution of the Internet: A Statistical Physics Approach* (Cambridge University Press, Cambridge, 2004); S. N. Dorogovtsev and J. F. F. Mendes, *Evolution of Networks: From Biological Nets to the Internet and WWW* (Oxford University Press, Oxford, 2003).
 - [2] V. Colizza, A. Barrat, M. Barthélemy, and A. Vespignani, *Proc. Nat. Acad. Sci. USA* **103**, 2015 (2006).
 - [3] J.-P. Onnela, J. Saramäki, J. Hyvönen, G. Szabo, D. Lazer, K. Kaski, J. Kertész, and A.-L. Barabási, *Proc. Nat. Acad. Sci. USA* **104**, 7332 (2007).
 - [4] S. Sreenivasan, R. Cohen, E. López, Z. Toroczkai, and H. E. Stanley, *Phys. Rev. E* **75**, 036105 (2007).
 - [5] S. S. Shen-Orr, R. Milo, S. Mangan, and U. Alon, *Nature Gen.* **31**, 64 (2002).
 - [6] R. Albert, H. Jeong, and A. L. Barabási, *Nature* **406**, 6794 (2000); **406**, 378 (2000).
 - [7] R. Cohen, K. Erez, D. ben-Avraham, and S. Havlin, *Phys. Rev. Lett.* **85**, 4626 (2000).
 - [8] Y. Chen, E. López, S. Havlin, and H. E. Stanley, *Phys. Rev. Lett.* **96**, 068702 (2006).
 - [9] D. J. Watts and S. H. Strogatz, *Nature* **393**, 440 (1998).

- [10] S. Fortunato, Phys. Rep. **486**, 75 (2010).
- [11] M. A. Porter, J. P. Onnela, P. J. Mucha, Notices of the AMS **56**, 1082 (2009).
- [12] *Social Network Analysis*, S. Wasserman and K. Faust (Cambridge University Press, Cambridge, 2005).
- [13] S. P. Borgatti and D. S. Halgin in *The Sage Handbook of Social Network Analysis* Carrington, P. and Scott, J. (eds) (Sage Publications Ltd, 2011).
- [14] P. Wang, K. Sharpe, G. L. Robins, and P. E. Pattison, Social Networks **31**, 12 (2009).
- [15] G. Ghoshal, V. Zlatić, G. Caldarelli, and M. E. J. Newman, Phys. Rev. E **79**, 066118 (2009).
- [16] *Hypergraphs, Volume 45: Combinatorics of Finite Sets* Claude Berge (North Holland, 1989).
- [17] C.M. Fortuin and P.W. Kasteleyn, Physica **57**, 536 (1972).
- [18] S. Yoon, A. V. Goltsev, S. N. Dorogovtsev, and J. F. F. Mendes, Phys. Rev. E **84**, 041144 (2011).
- [19] M. E. J. Newman, Phys. Rev. Lett. **103** 058701 (2009).
- [20] J. C. Miller, Phys. Rev. E **80**, 020901(R) (2009).
- [21] The relation between bipartite graphs and hypergraphs is well known, and involves the duality transformation of assigning to every hyperedge an “affiliation” node in a bipartite graph.
- [22] J. Park and M. E. J. Newman, Phys. Rev. E **70**, 066117 (2004).
- [23] S. Bradde and G. Bianconi, J. Phys. A: Math. Theor. **42**, 195007 (2009); S. Bradde and G. Bianconi, J. Stat. Phys. P07028 (2009).
- [24] A. Engel, R. Monasson, and A. K. Hartmann, J. Stat. Phys. **117**, 387 (2004).
- [25] *Introduction to Combinatorial Analysis*, J. Riordan (John Wiley & Sons., New York (1958)).
- [26] *Advanced Mathematical Methods for Scientists and Engineers*, C. M. Bender and S. A. Orszag (Springer).

Set notation	Explanation	Type of element	Size
$\mathbf{T}_{ij}(N, r)$	Hyperedges of complete hypergraph simultaneously visiting i and j	hyperedge	$\binom{N-2}{r-2}$
$O_{ij}(\sigma)$	Hyperedges of configuration σ simultaneously visiting i and j	hyperedge	o_{ij}
\mathbf{O}_{ij}	Collection of all possible sets $O_{ij}(\sigma)$	Set of cardinality o_{ij} of hyperedges	$\binom{N-2}{o_{ij}}$

TABLE I: Notation used for calculation of $\phi_{ij}(o_{ij}, \mathbf{p})$. The complement sets $\overline{O}_{ij}(\sigma)$ are with respect to $\mathbf{T}_{ij}(N, r)$.

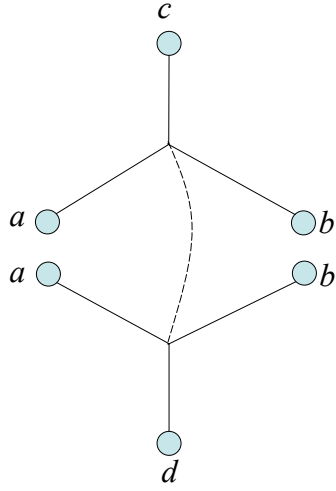
Set notation	Explanation	Type of element	Size
$\mathbf{T}_i(N, r)$	Hyperedges of complete hypergraph visiting i	hyperedge	$\binom{N-1}{r-1}$
$\lambda_i(\sigma)$	Hyperedges of configuration σ visiting i	hyperedge	ℓ_i
$\mathbf{\Lambda}_i$	Collection of all possible sets $\lambda_i(\sigma)$	Set of cardinality ℓ_i of hyperedges	$\binom{N-1}{\ell_i}$

TABLE II: Notation used for calculations of $\zeta_i(\ell_i, \mathbf{p})$ and $\xi_i(s_i, \mathbf{p})$ in the \mathcal{P}_a projection. The complement sets $\overline{\lambda}_i(\sigma)$ are with respect to $\mathbf{T}_i(N, r)$.

Set notation	Explanation	Type of element	Size
$K_i^{(\ell_i)}(\boldsymbol{\sigma})$	Hyperedges in configuration $\boldsymbol{\sigma}$ visiting i plus k_i other nodes	hyperedge	ℓ_i
$I_i^{(\rho(k_i), \ell_i)}(\boldsymbol{\sigma})$	Hyperedges in configuration $\boldsymbol{\sigma}$ visiting i plus the k_i nodes in set $\rho(k_i)$	hyperedge	ℓ_i
$\rho(k_i)$	Choice of k_i nodes (plus i) in $\boldsymbol{\sigma}$ connected to i via ℓ_i hyperedges	node	k_i
$\mathbf{K}_i(\ell_i)$	Collection of all possible sets $K_i^{(\ell_i)}(\boldsymbol{\sigma})$	Set of cardinality ℓ_i of hyperedges	$\binom{N-1}{k_i} Q_{r-1}(k_i, \ell_i)$
$\mathbf{I}_i(\rho(k_i), \ell_i)$	Collection of all possible sets $I_i^{(\rho(k_i), \ell_i)}(\boldsymbol{\sigma})$	Set of cardinality ℓ_i of hyperedges	$Q_{r-1}(k_i, \ell_i)$
$\mathbf{R}_i(N, k_i)$	Collection of all possible sets $\rho(k_i)$	Set of cardinality k_i of nodes	$\binom{N-1}{k_i}$

TABLE III: Notation used for calculation of $\psi_i(k_i, \mathbf{p})$. The complement sets $\overline{K}_i^{(\ell_i)}(\boldsymbol{\sigma})$ and $\overline{I}_i^{(\rho(k_i), \ell_i)}(\boldsymbol{\sigma})$ are with respect to $\mathbf{T}_i(N, r)$.

Hypergraph:
 $\sigma_{abc}=1, \sigma_{abd}=1$, all other $\sigma_{ijg}=0$



Projected network:

$$w_{ij} = \sum_g \sigma_{ijg}$$

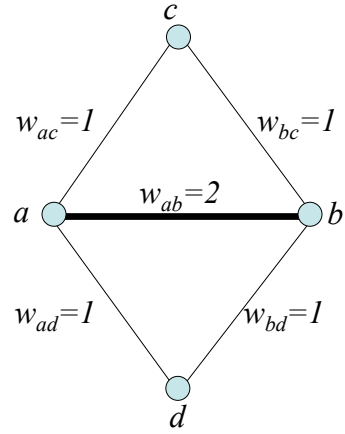


FIG. 1: Illustration for the projection \mathcal{P}_a from hypergraphs to networks. On the left, a hypergraph is composed of a multitude of hyperedges that exist when $\sigma = 1$, and do not when $\sigma = 0$. The projected network (right) has a link between all nodes that belong to the same hyperedge, and the weight of the link is the number of hyperedges that share the same pair of nodes.

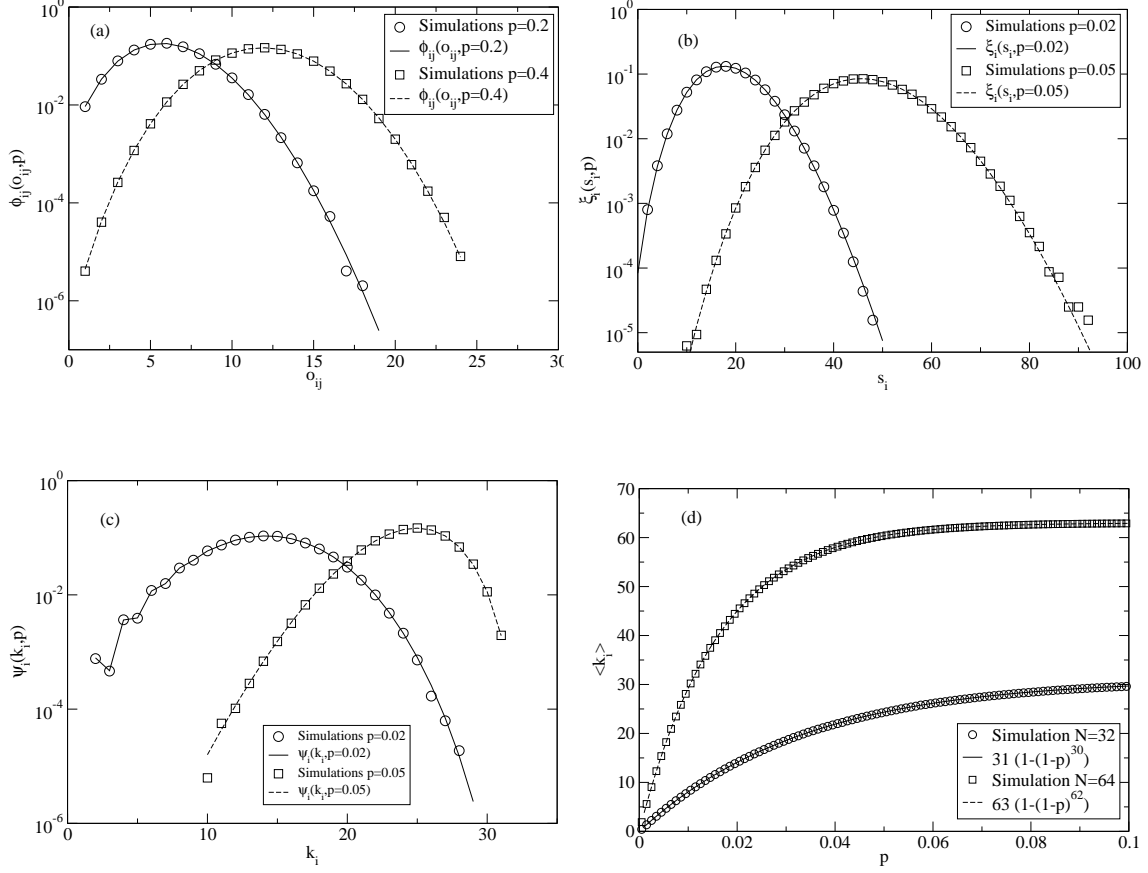


FIG. 2: Comparison between theoretical distributions (lines) and simulations (symbols) for distributions of homogeneous projected networks for $N = 32$ and $r = 3$: (a) $\phi_{ij}(o_{ij}, p)$ from Eq. (21) for $N = 32$ and corresponding simulations (\bigcirc for $p = 0.2$ and \square for $p = 0.4$); (b) $\xi_i(s_i, p)$ from Eq. (27) for $N = 32$ and corresponding simulation (\bigcirc for $p = 0.02$ and \square for $p = 0.05$); (c) $\psi_i(k_i, p)$ from Eq. (36) for $N = 32$ and corresponding simulations (\bigcirc for $p = 0.02$ and \square for $p = 0.05$). (d) Average degree $\langle k_i \rangle$ as a function of p in homogeneous networks from Eq. (38) and from simulations (\bigcirc for $N = 32$ and \square for $N = 64$).

Sets: $\rho(k_i=3)=\{a,b,c,i\}$
 $\mathbf{I}_i(\rho(k_i=3),\ell_i=2)=\{\{(a,b,i),(b,c,i)\},\{(a,b,i),(a,c,i)\},\{(a,c,i),(b,c,i)\}\}$
 $\mathbf{I}_i(\rho(k_i=3),\ell_i=3)=\{\{(a,b,i),(a,c,i),(b,c,i)\}\}$

Configurations:

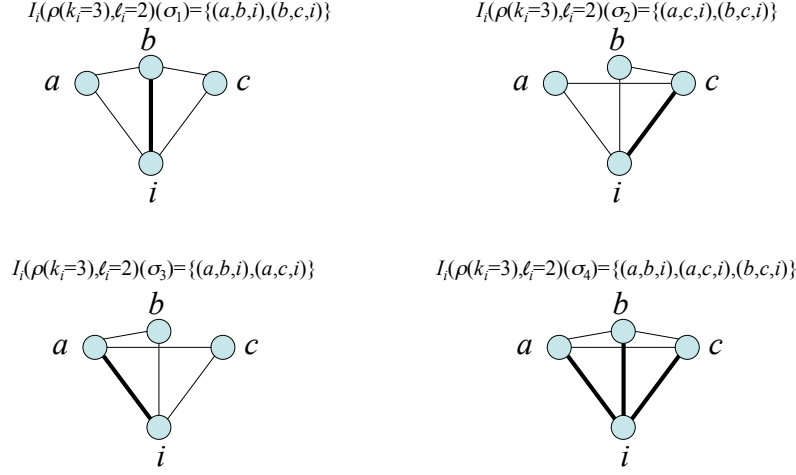


FIG. 3: Illustration ($r = 3$) of the emergence of degree k_i as a consequence of various possible hyperedge configurations. The figure also illustrates $Q_{r-1}(k_i, \ell_i)$. There are 4 possible ways in which i can be connected to nodes $\{a, b, c\}$, each case corresponding to one of the configurations shown above ($\sigma_1, \sigma_2, \sigma_3, \sigma_4$) in the projected network. The sets $\mathbf{I}_i(\rho(k_i), \ell_i)$ are defined for both $\ell_i = 2$ and 3, the only two possible cases. Note also that if one focuses only on the nodes $\{a, b, c\}$ ignoring i , all configurations can be mapped to the construction of all possible cliques of size 2 of these nodes, generating $Q_{r-1=2}(k_i = 3, \ell_i = 2) = 3$ and $Q_{r-1=2}(k_i = 3, \ell_i = 3) = 1$. The fact that all configurations are globally connected is an accident due to the small value of $k_i = 3$, but in general nodes simply need to belong to ℓ_i cliques of size $r - 1$. Finally, note the thickness of links, representative of o_{ij} .

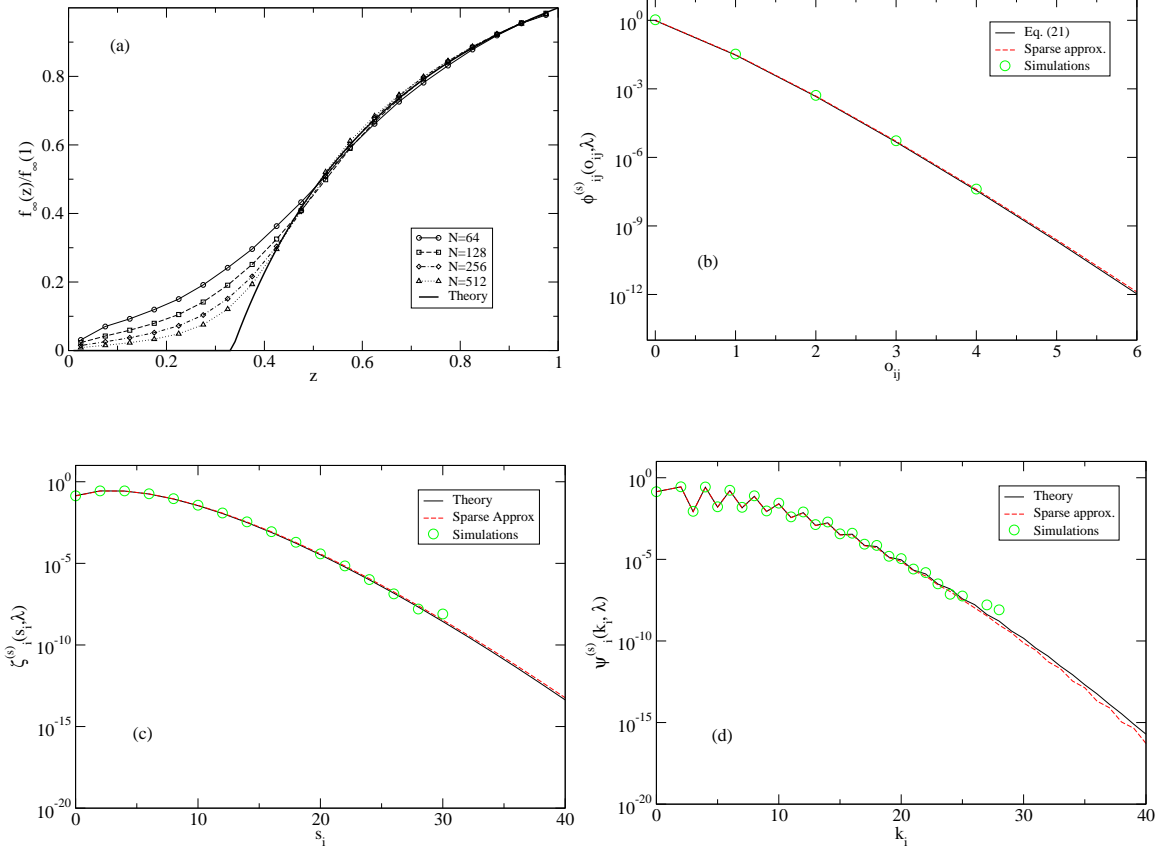


FIG. 4: (color online) Percolation limit for the ensemble of $\langle s \rangle = \lambda_{\max}$: (a) $f(z)/f(z=1)$ vs. z ($\lambda_{\max} = 3.0$) from Eq. (50) (line) and simulations of $N = 64$ (\circ), $N = 128$ (\square), $N = 256$ (\diamond) and $N = 512$ (\triangle). As the system size increases, the theoretical solution is approached. Projected network properties (\mathcal{P}_a for s_i) for $N = 128$ in the ensemble of $\langle s \rangle = \lambda = 4.0$ predicted by theory (line), their respective sparse approximations (dashed line) and simulations (\circ): (b) $\phi_{ij}^{(s)}(o_{ij}, \lambda)$, (c) $\xi_i^{(s)}(s_i, \lambda)$, and (d) $\psi_i^{(s)}(k_i, \lambda)$.